International PhD Thesis

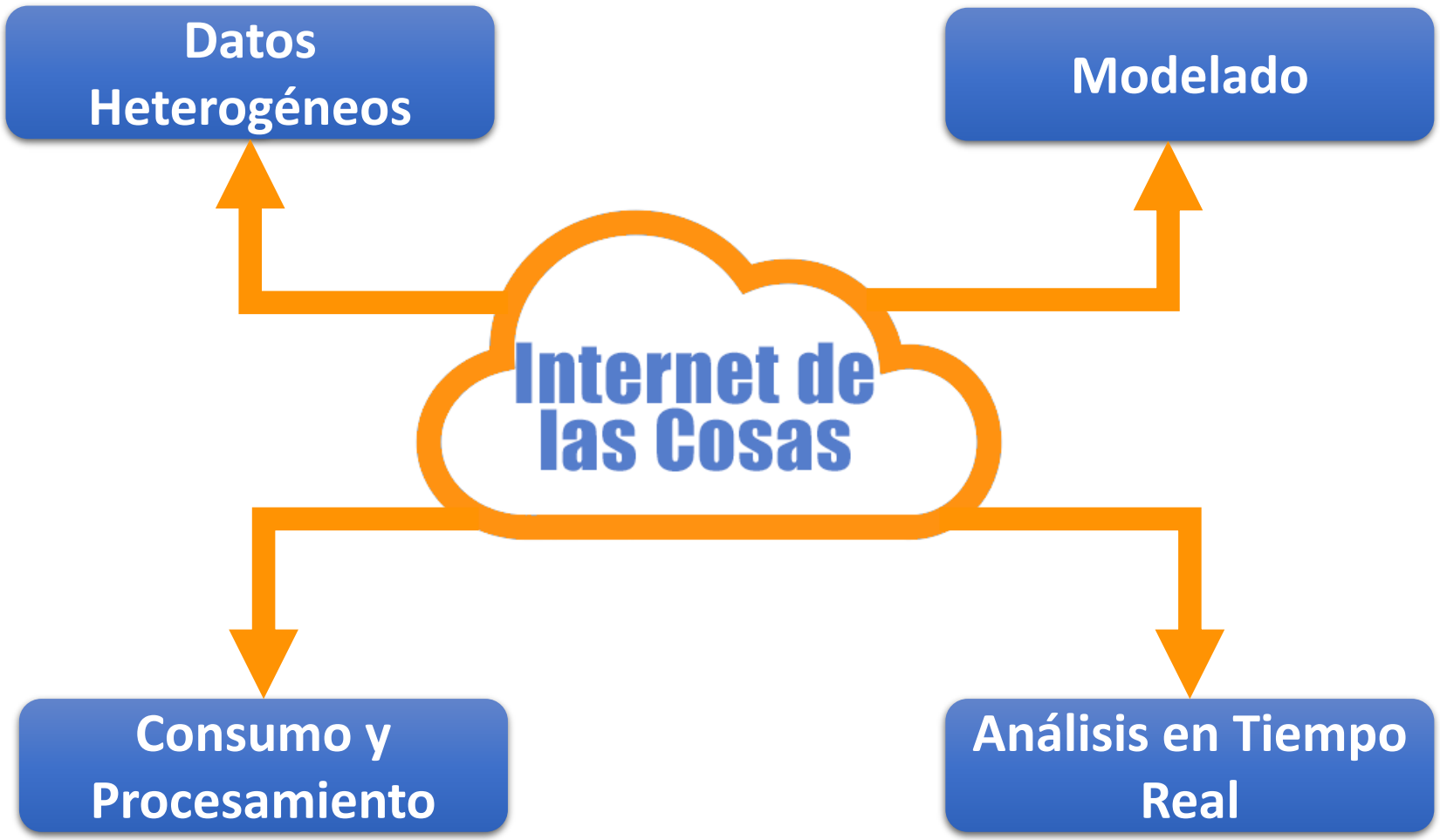
Processing Heterogeneous Data in the Internet of Things

David Corral Plaza

Supervisors:
Guadalupe Ortiz Bellot, Inmaculada Medina Bulo

University of Cádiz, 8th March 2021

Resumen


Datos Heterogéneos
Modelado
Internet de las Cosas
Consumo y Procesamiento
Análisis en Tiempo Real

Abstract

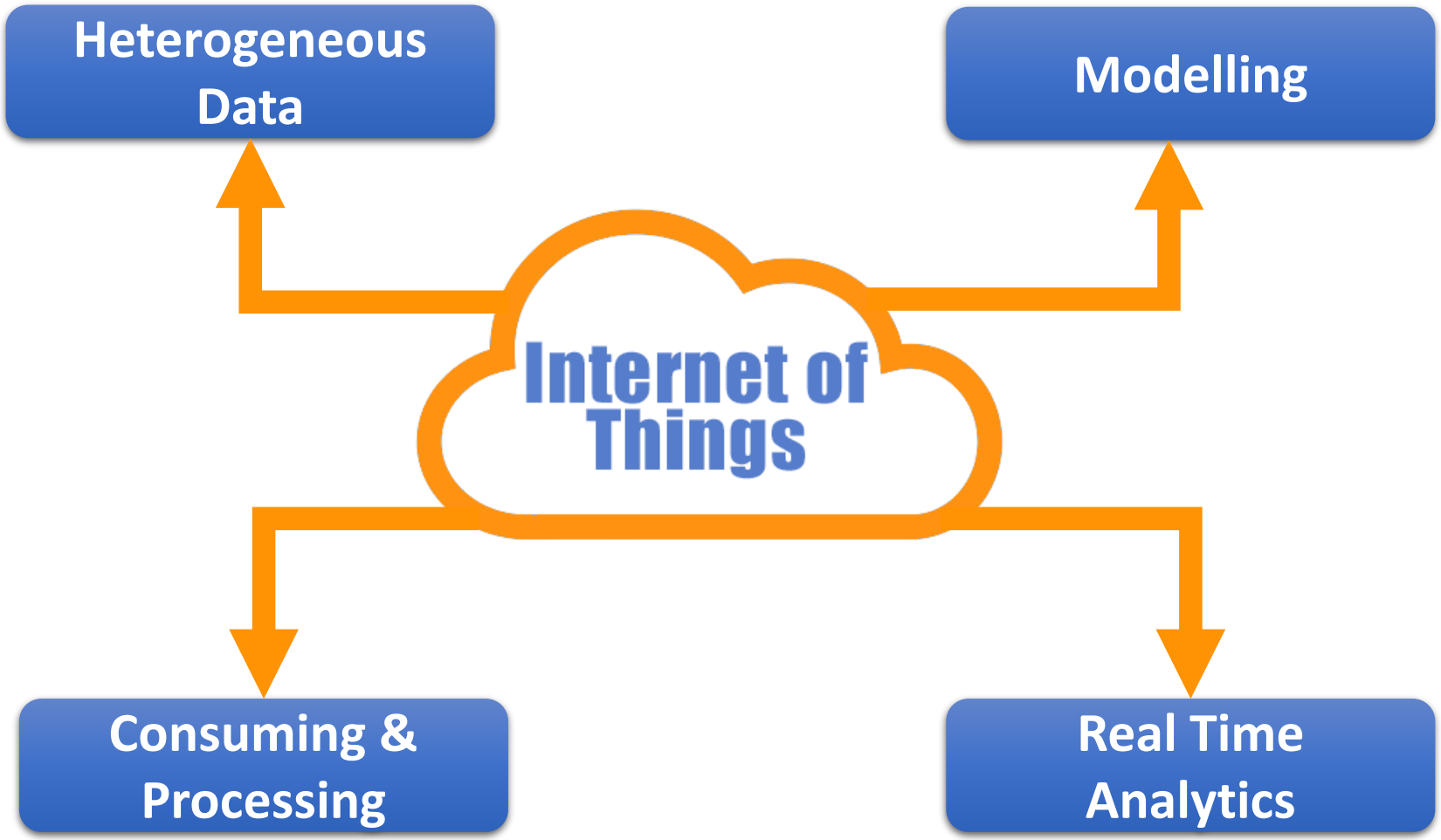

Heterogeneous Data
Modelling
Internet of Things
Consuming & Processing
Real Time Analytics

Table of Contents

Table of Contents

Motivation

Increasing number of **Internet of Things (IoT) devices** and growing IoT market share.

Dealing with **heterogeneity**, **speed**, and **volume** in IoT researching.

Analyse **large heterogeneous IoT datasets** in real time, results in a **competitive advantage**.

Domain experts as the main users.

Issues

Heterogeneity

- We must be able to analyse data regardless of their structure.

Real time analytics

- We must be able to analyse and react to these data as soon as possible.

Friendly and intuitive usage

- We must be able to provide these functionalities to any kind of users.

Goals

Providing users with the capability of modelling, consuming, processing, analysing, and detecting situations of interest in IoT domains through an intuitive and friendly set of tools.

Design and implementation of an architecture for processing heterogeneous data.

Design and implementation of a friendly and intuitive graphical editor.

Combine and integrate both, the architecture for data processing and the graphical editor for modelling.

Exhaustive evaluation of the proposed system.

Table of Contents

Technological Background

Data Serialization Systems

Complex Event Processing

Stream Processing

Model-Driven Development

Data Serialization Systems

Data Serialization Systems (DSS) allows us to convert the data into lighter formats.

Powerful to **process**, **store**, and **transport** large amounts of data.

Apache Avro and their Schemas.

Avro Generic Records in the homogenisation task.

Complex Event Processing

Complex Event Processing (CEP) is a well-stablish stream analytics technology.

No need to store data, **we store patterns**.

Real time feedback means improved **decision making**.

Esper for CEP analytics.

Stream Processing

Stream Processing (SP) is a computer paradigm to process streams of unbounded data in real time.

Main benefits: **Unlimited** streams of data, Event-Based Architectures (**EDA**), **Low latency** strategy, or good **Scalability**.

Apache Kafka Streams as SP.

Model-Driven Development

Model-Driven Development (MDD) focuses on the main aspects.

Represented using a Domain-Specific Modelling Language (**DSML**).

A **DSML** is composed of a **metamodel**, **restrictions**, **specific syntax**, and **transformations**.

Main benefits: improved **productivity**, **communication with domain experts**, **adaptability** and **goals.**

Table of Contents

Existing Approaches

Stream analytics as the main solution.

Homogenisation in the sources or **Normalisation** of the heterogeneous data in batches.

Performance and **changes at runtime**.

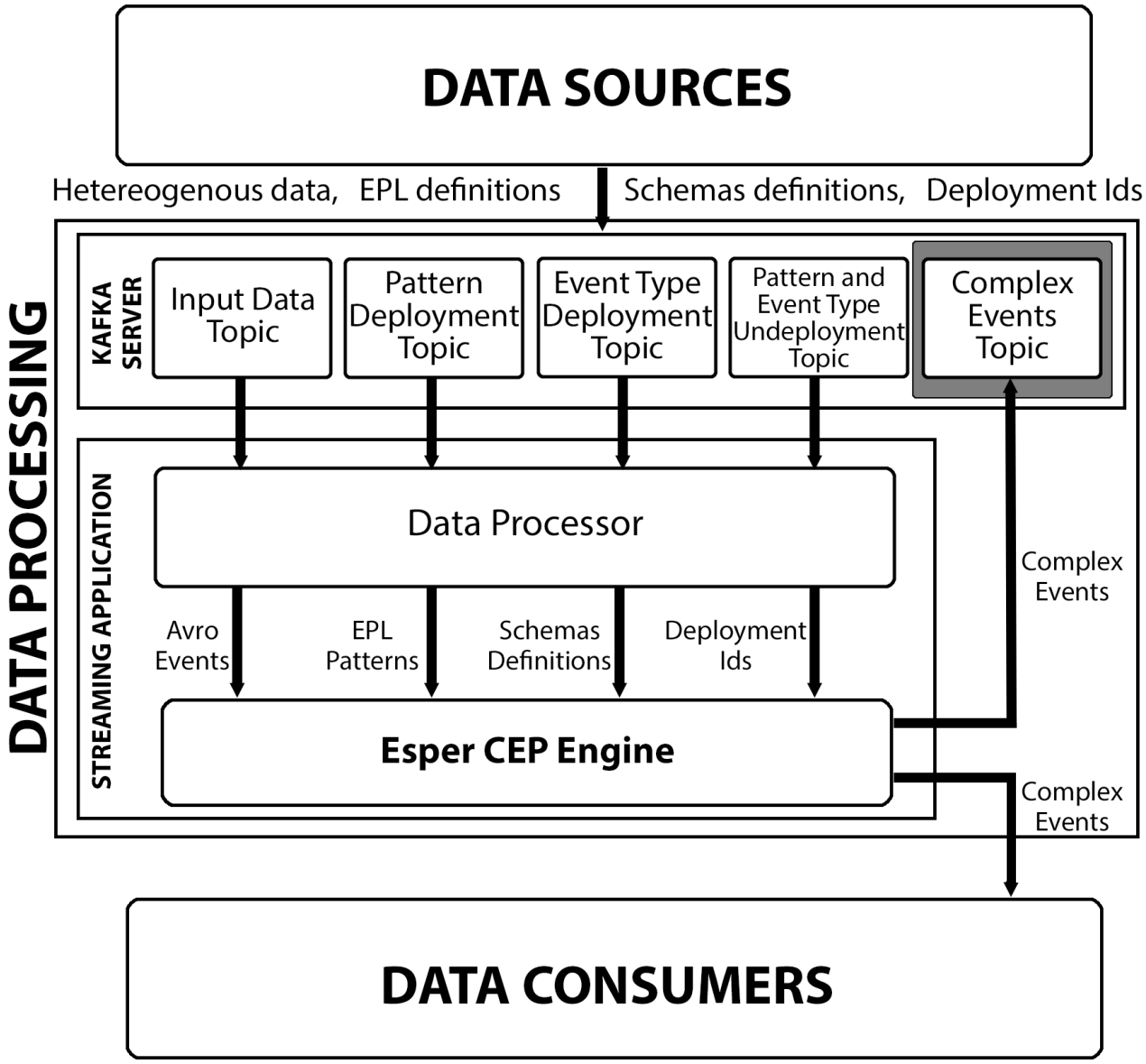
No approaches **integrating MDD** with **SP**.

Table of Contents

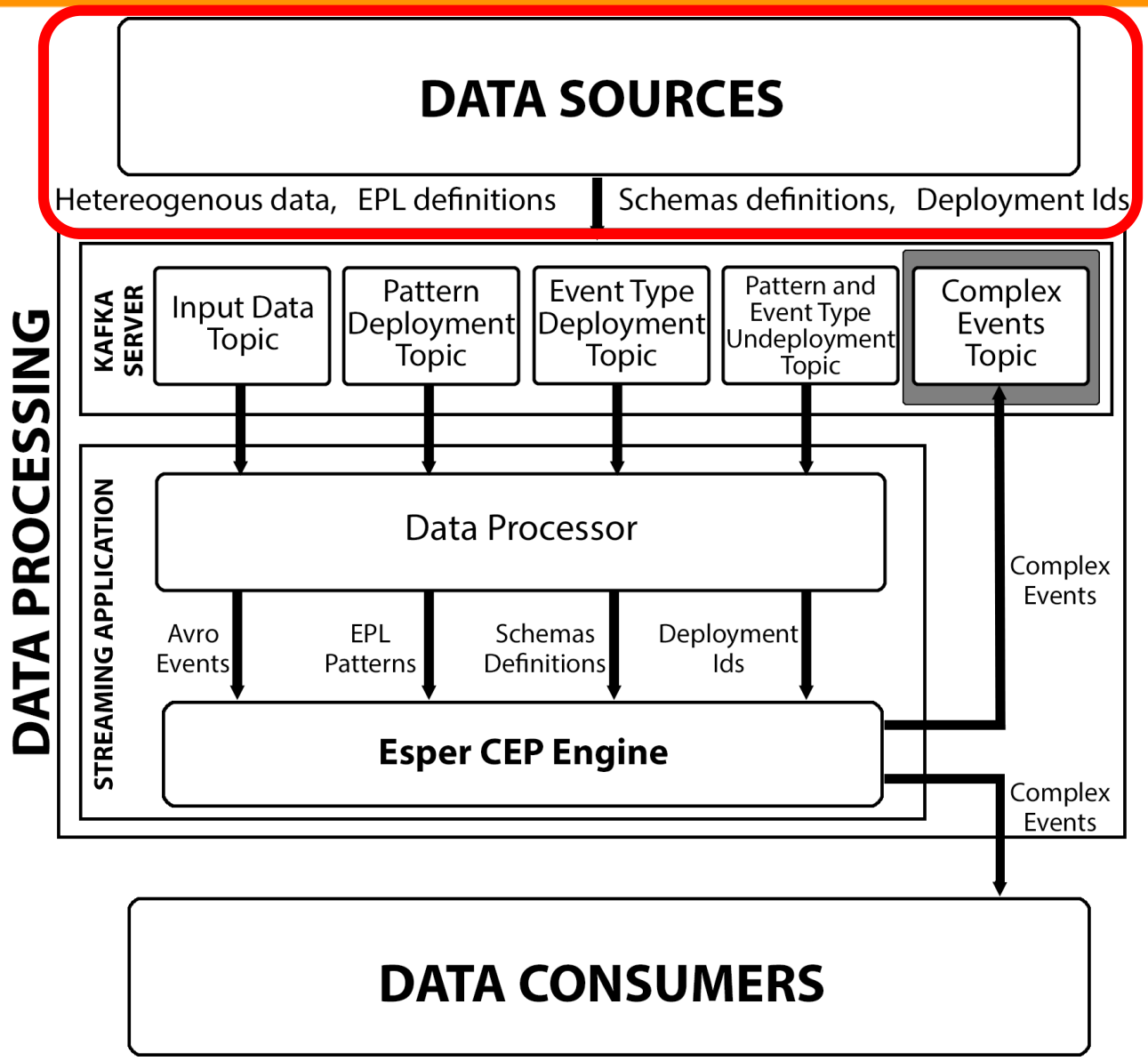
Key Features

- (F1) Scalability
- (F2) Low latency
- (F3) Data Analytics
- (F4) Data Storage
- (F5) Heterogeneous data
- (F6) Predictive Tools
- (F7) Real Time
- (F8) Performance Evaluation

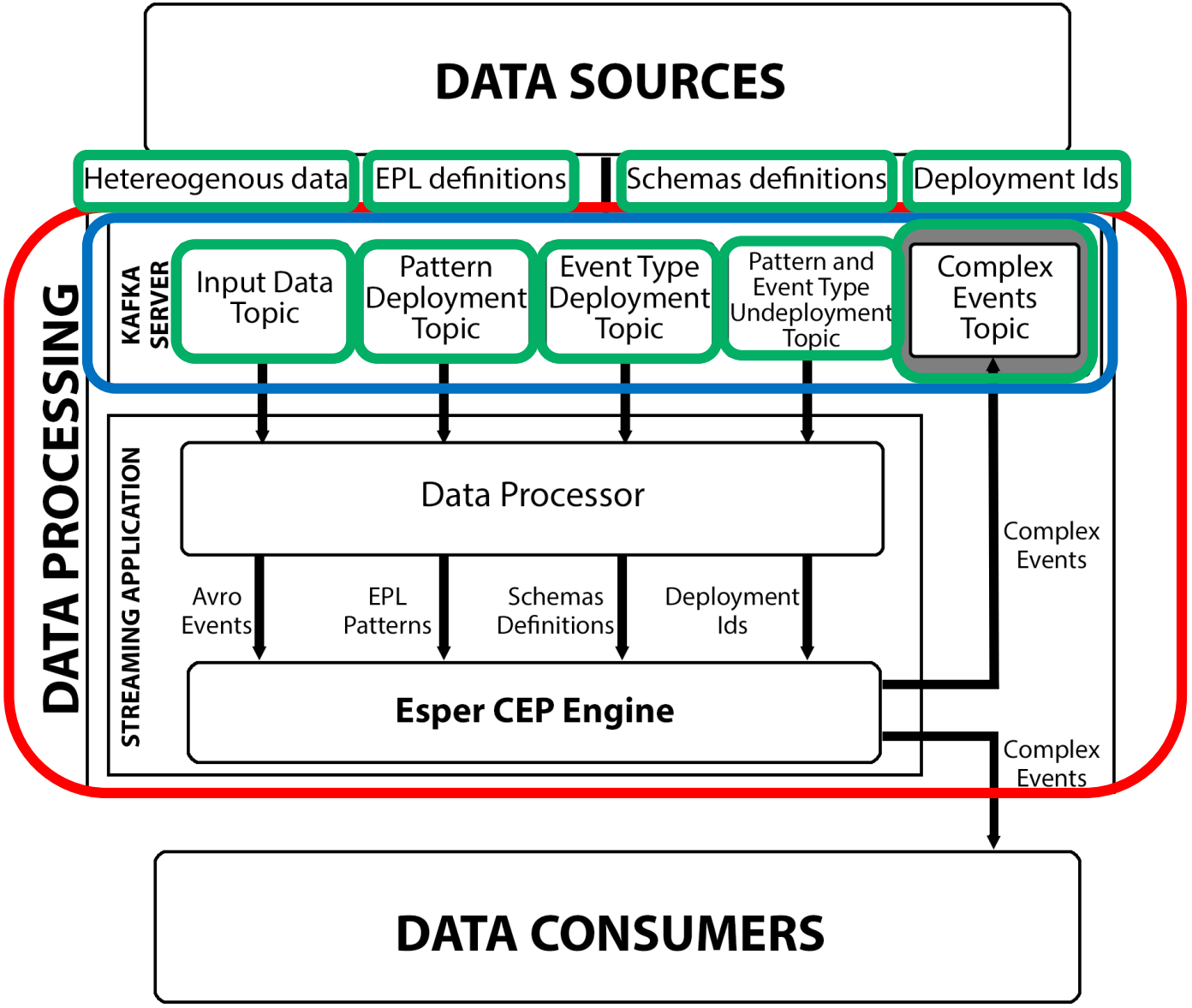
Stream Processing Architecture



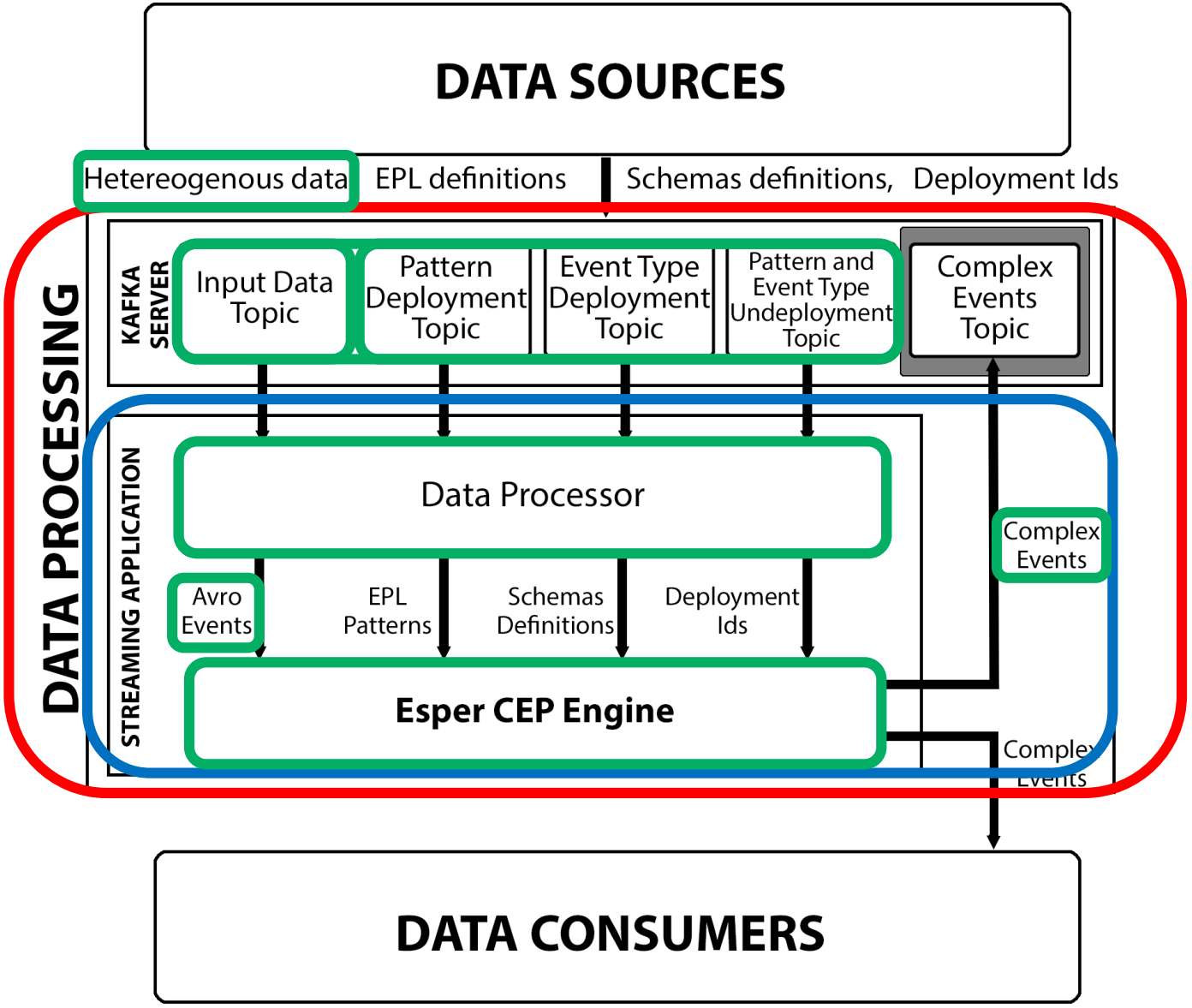
Stream Processing Architecture



Stream Processing Architecture



Stream Processing Architecture



Stream Processing Architecture

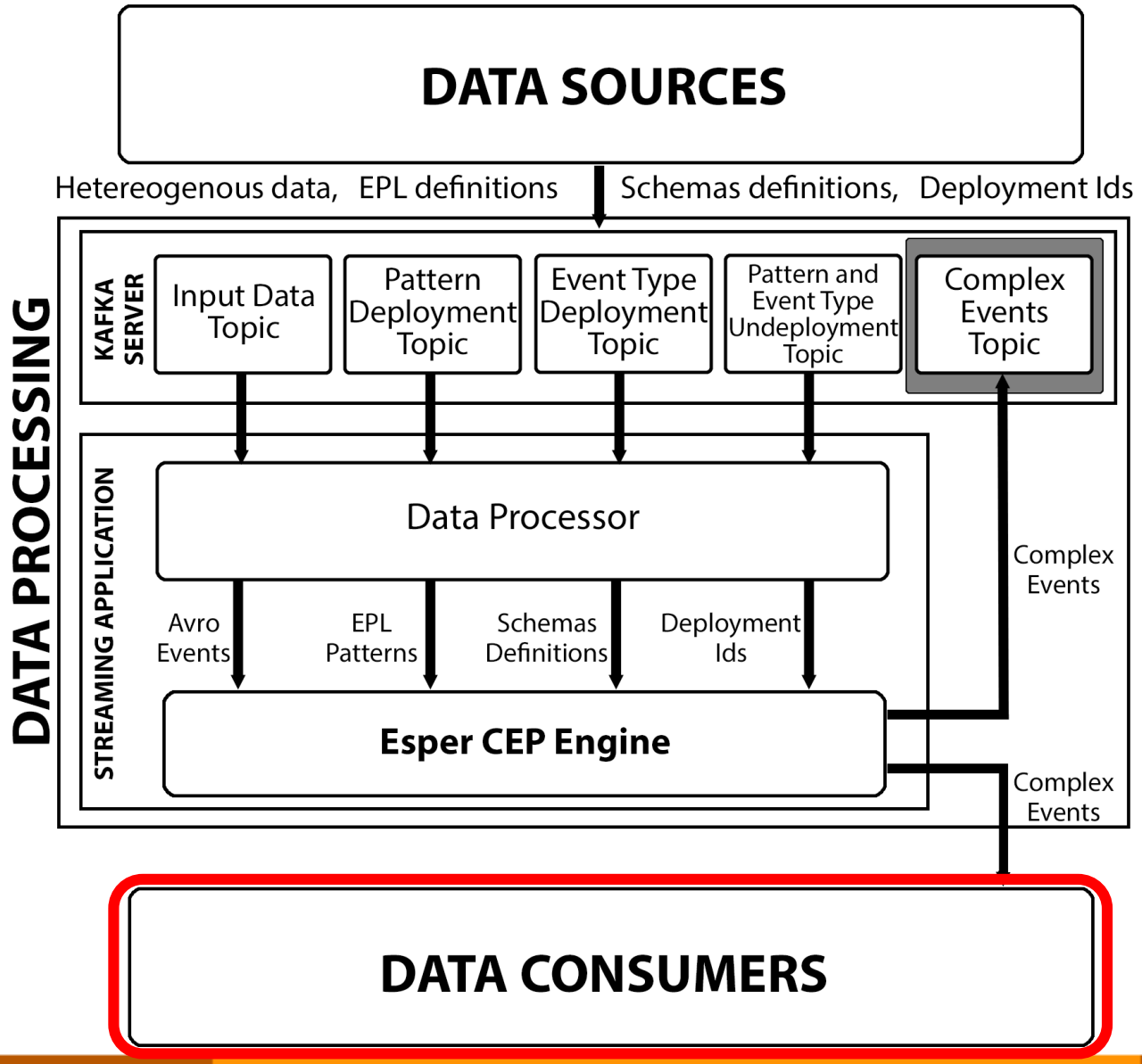


Table of Contents

Motivation

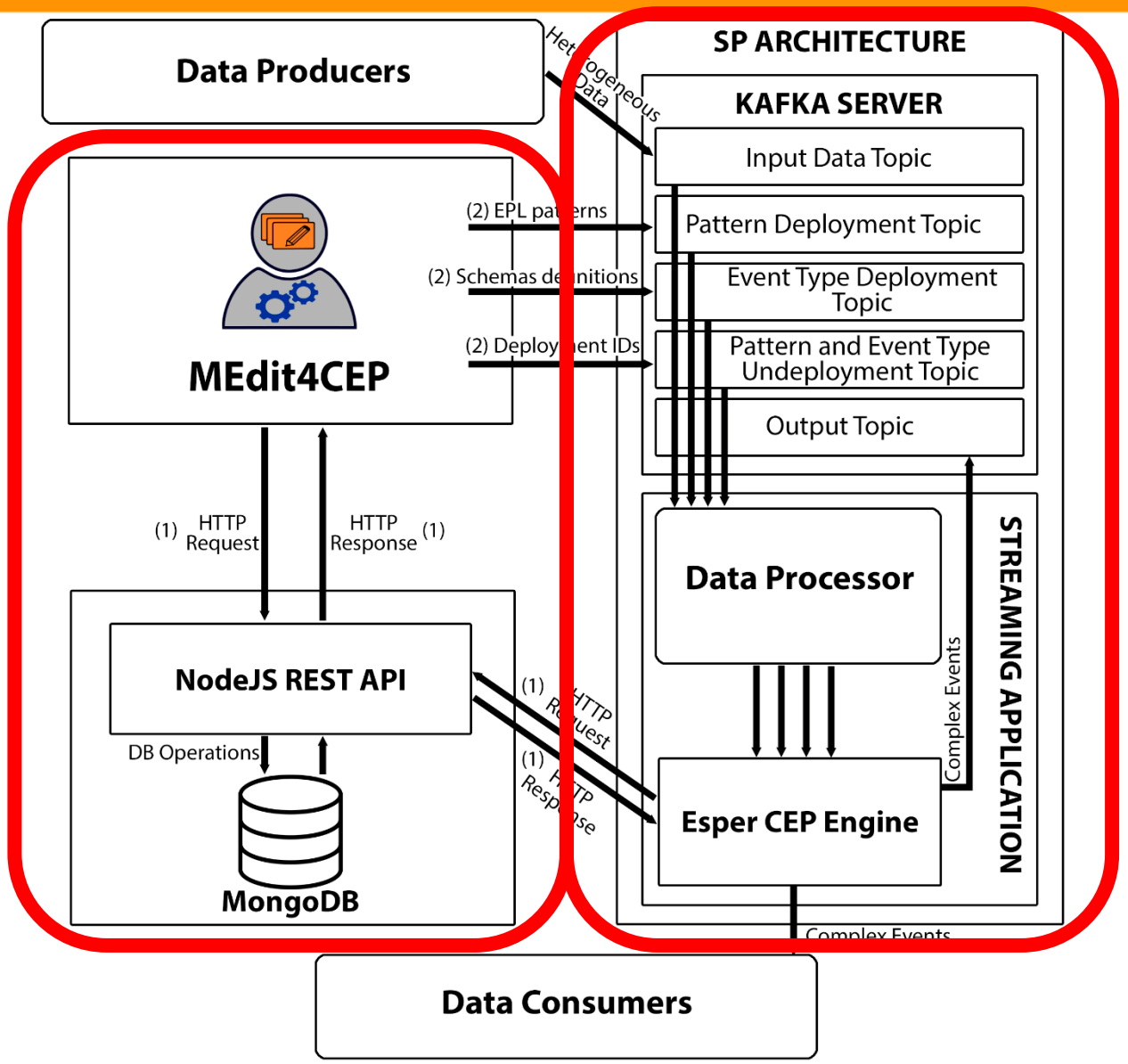
A SP architecture for processing heterogeneous data.

MEdit4CEP: a MDD graphical editor for modelling CEP domains and situations of interest.

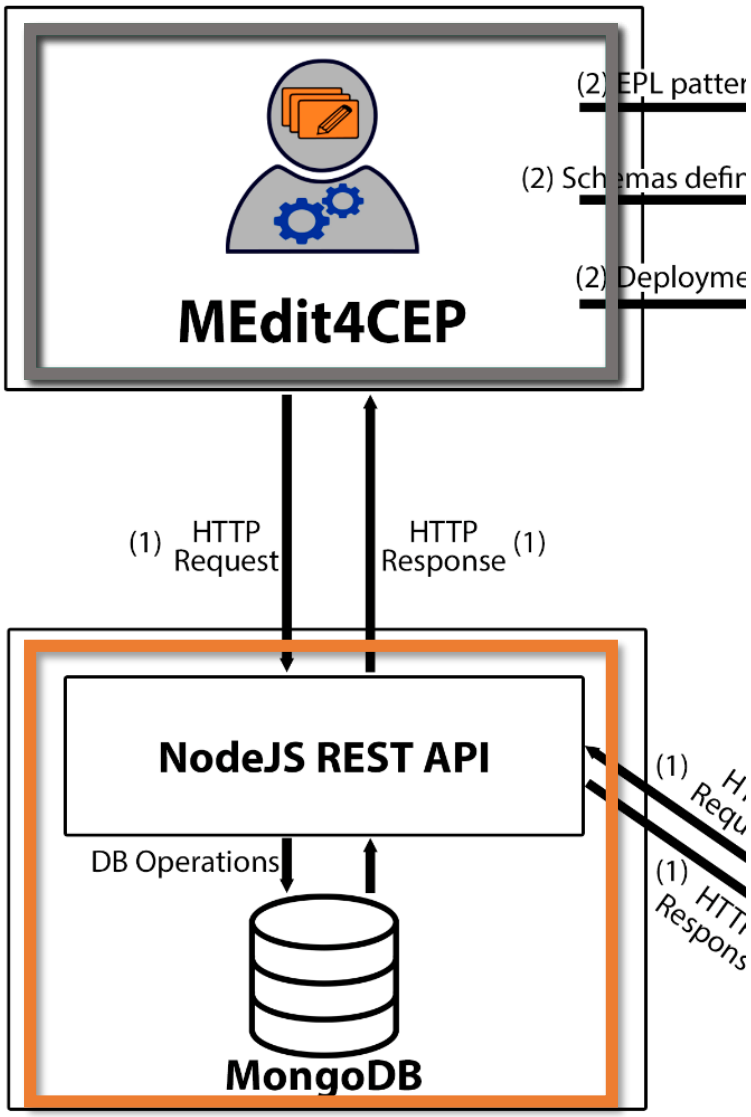
The integration of both, brings these complex technologies to **any kind of users**.

A new set of tools, named **MEdit4CEP-SP**.

MEdit4CEP-SP



MEdit4CEP-SP



RESTful API & NoSQL database for managing the **persistence**.

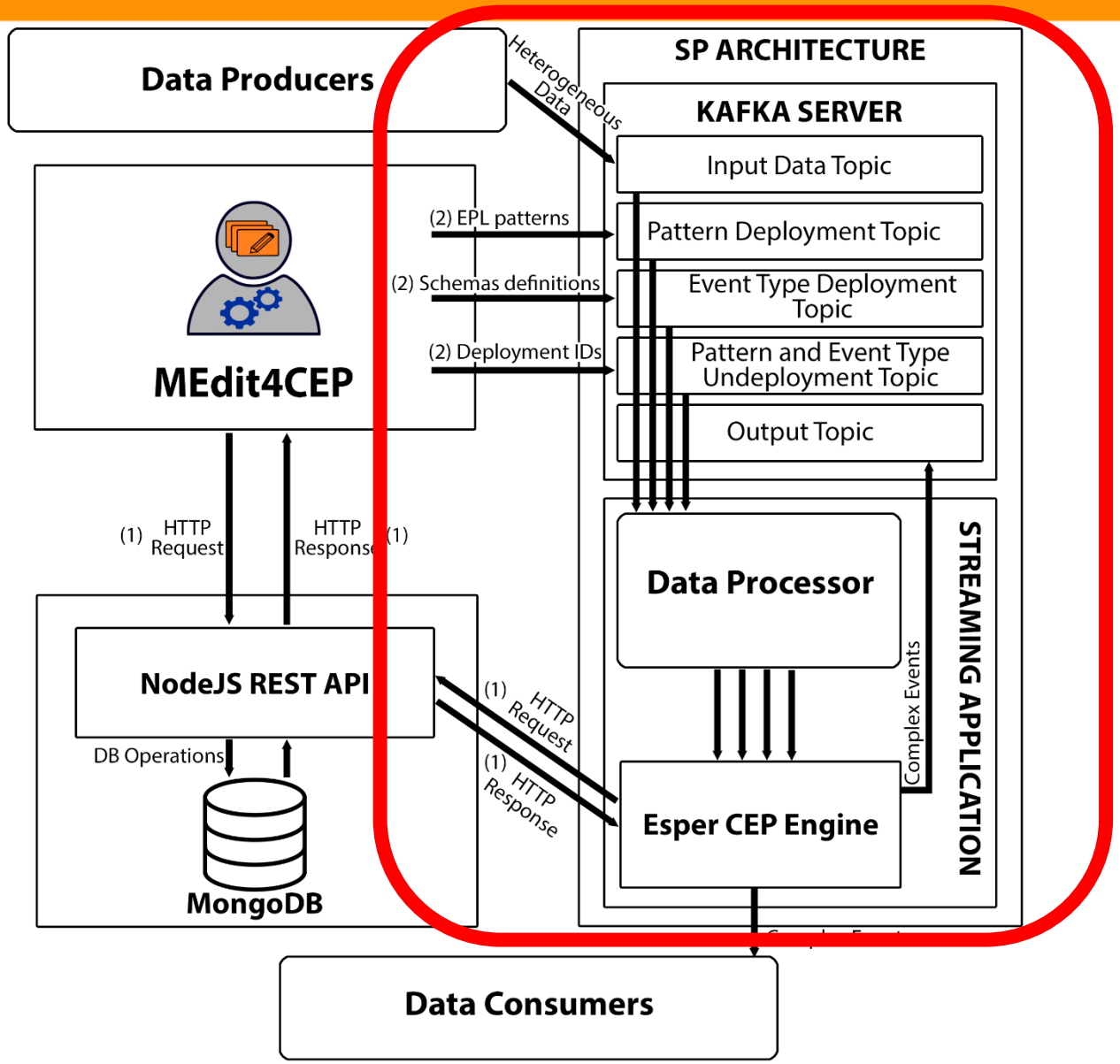
Avro Schema and **EPL** Generation.

Improving **CEP domain Autodetection**.

Improving **model change detections**.

These changes aim to **integrate** MEdit4CEP and the SP architecture.

MEdit4CEP-SP



MEdit4CEP-SP

Deployment of event types and event patterns at **runtime**.

Autodetection of event types from the SP architecture.

Detection of changes on deployed event types and event patterns.

Removal and **update** of event types and event patterns at **runtime**.

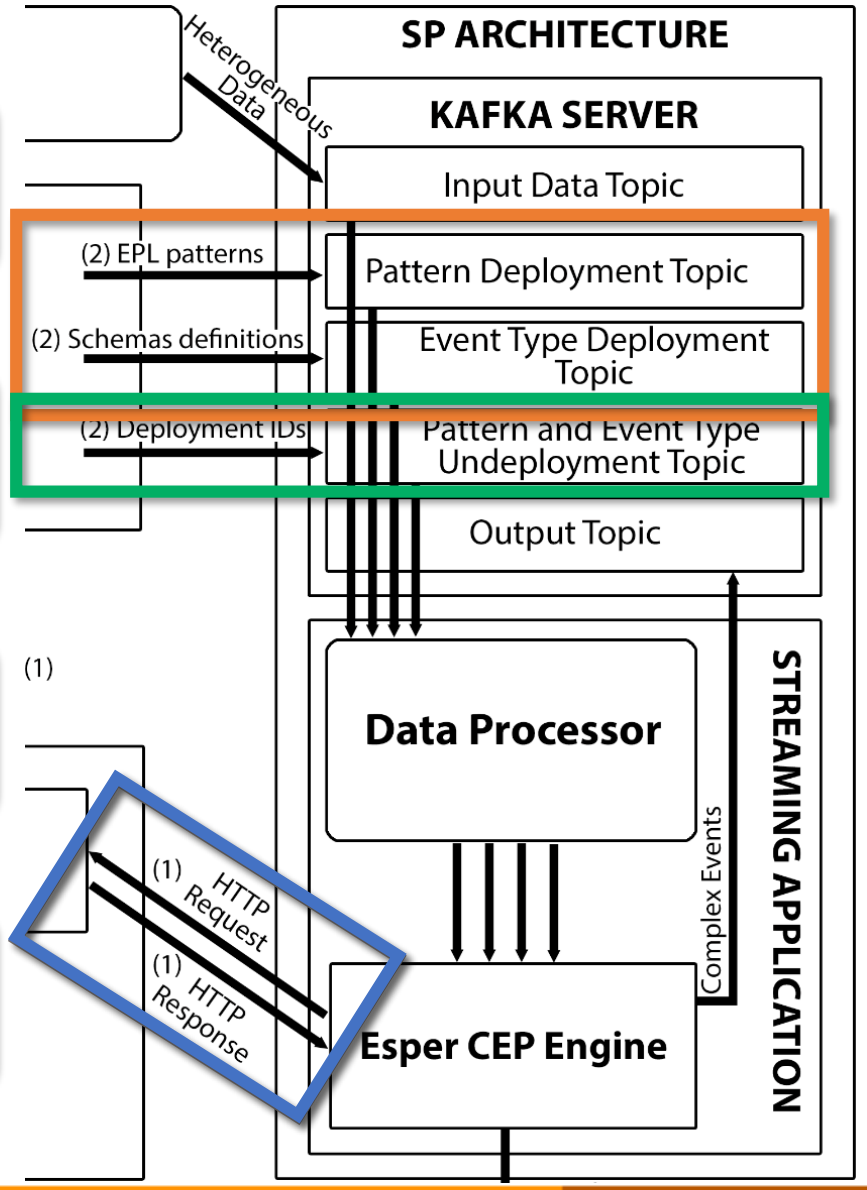


Table of Contents

Methodology

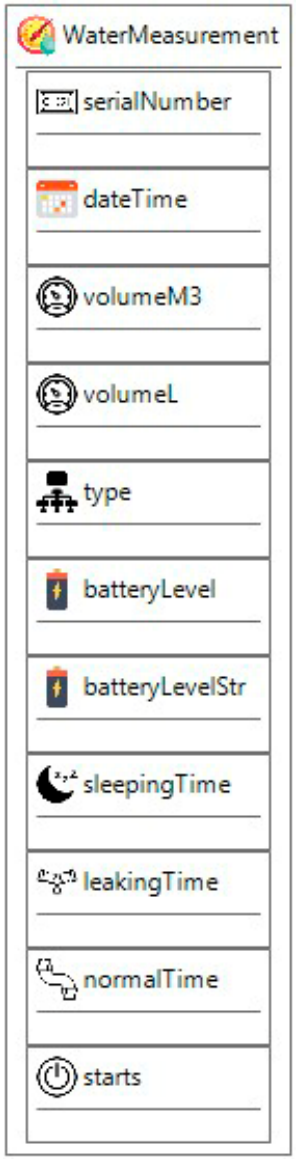
Two case studies: **Smart water management** and **Human activity recognition**.

Aim to show the **applicability** of the proposed solution.

In **both** case studies, the procedure is:

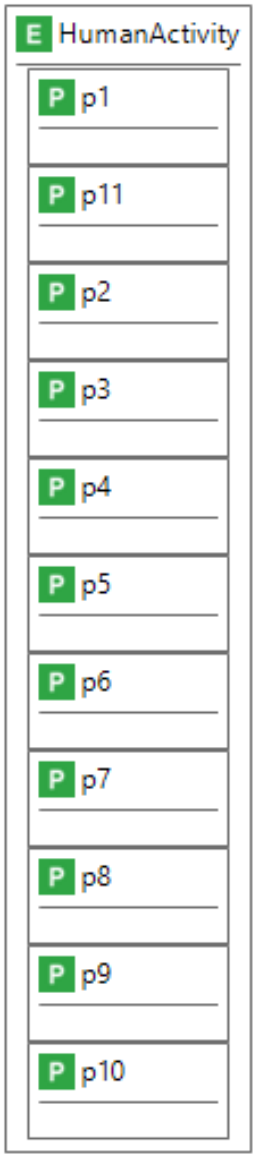
CEP Domain Definition	Situations of Interest Definition	Heterogeneous Data Simulation	Situations of Interest Detection

Step 1: CEP Domain Definition



Smart Water Domain

- Defined by the **user** using the editor of **MEdit4CEP-SP**.
- Transformed into **Avro Schema** and into **Avro EPL**.
- Deployed into the **SP architecture** and stored in the **NoSQL database**.



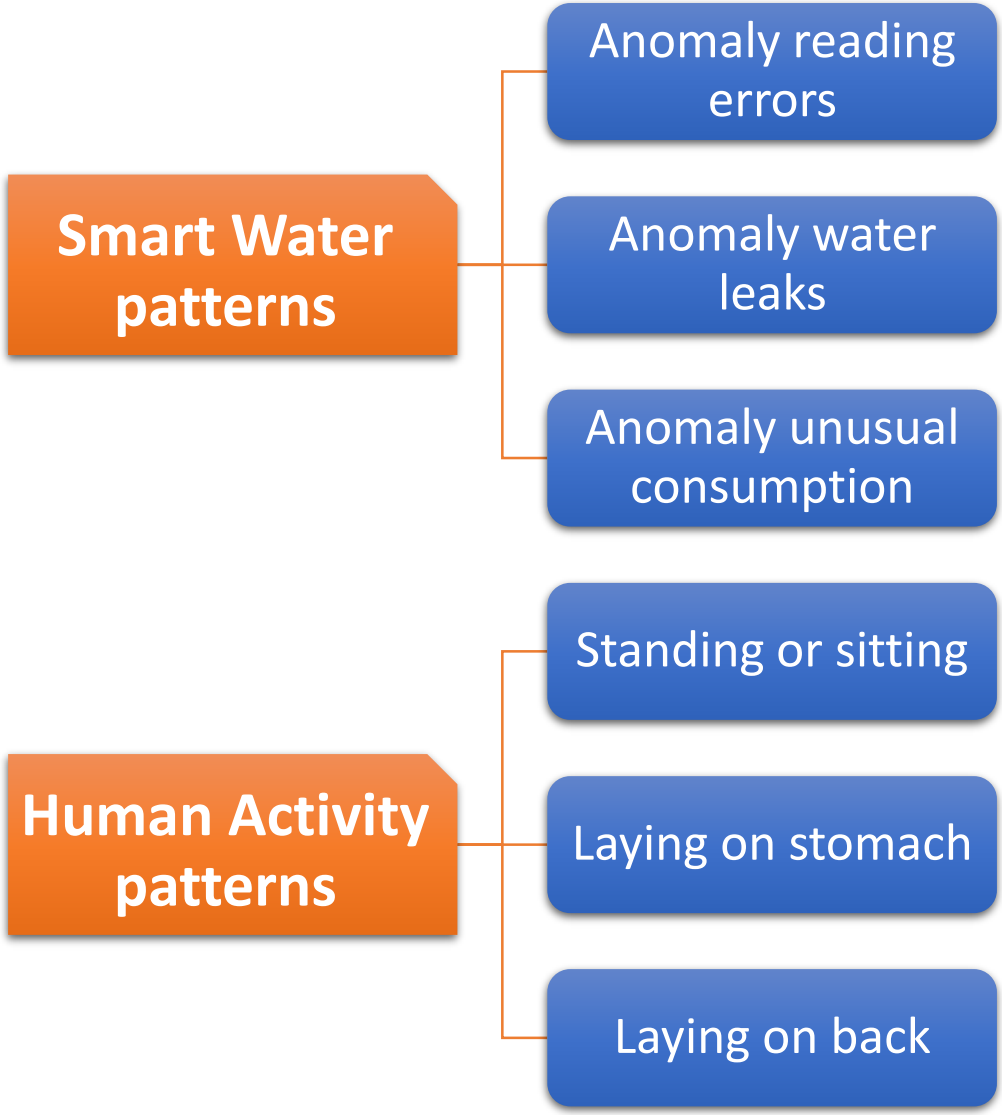
Human Activity Domain

- Inferred by the **SP architecture**.
- Transformed into **Avro Schema** and sent to the **API**.
- Retrieved in the editor of **MEdit4CEP-SP** using the **CEP Domain Autodetection** option.

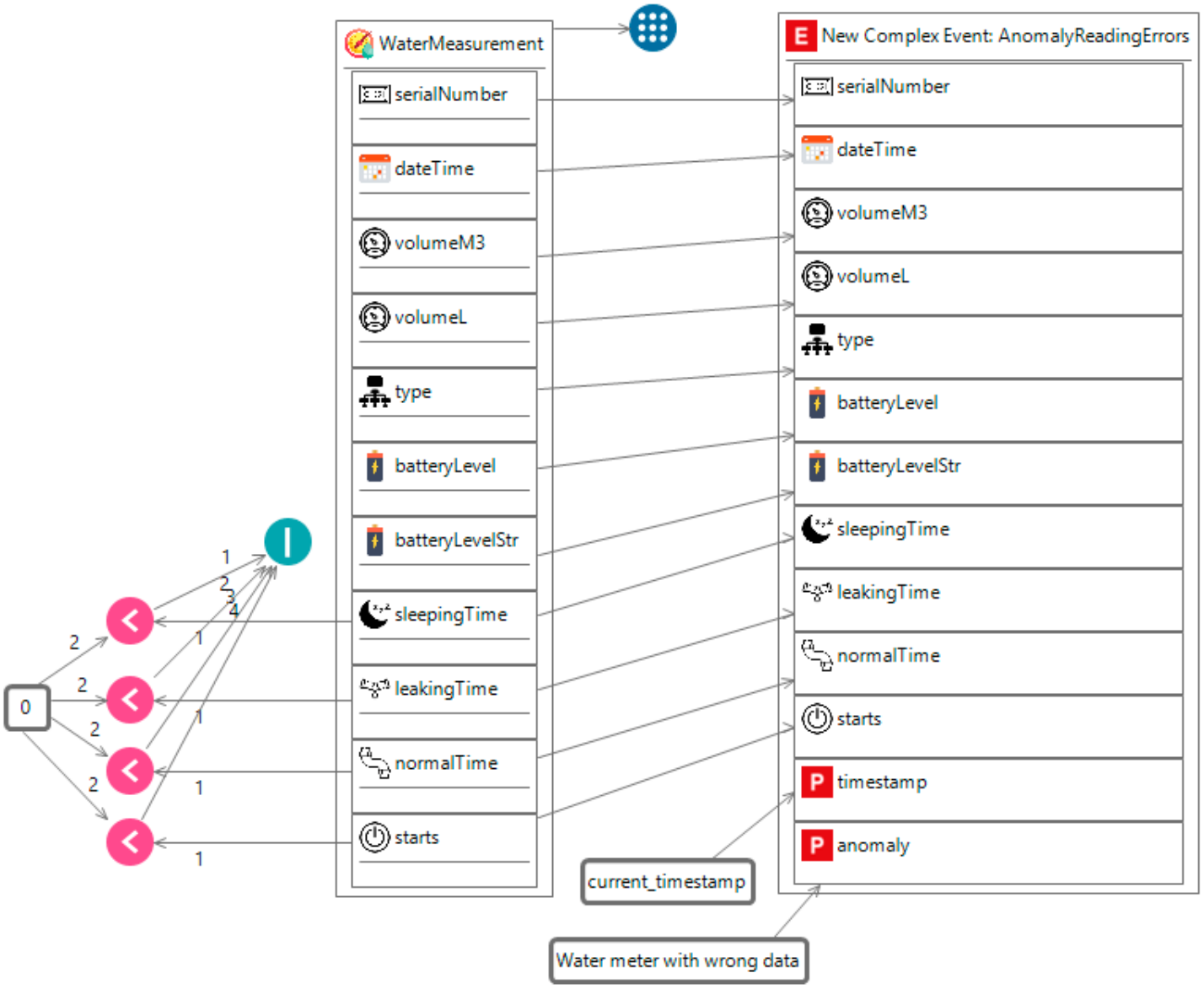
Step 2: Situation of Interest Definition

Modelled, validated, and transformed into EPL code using the editor of **MEdit4CEP-SP**.

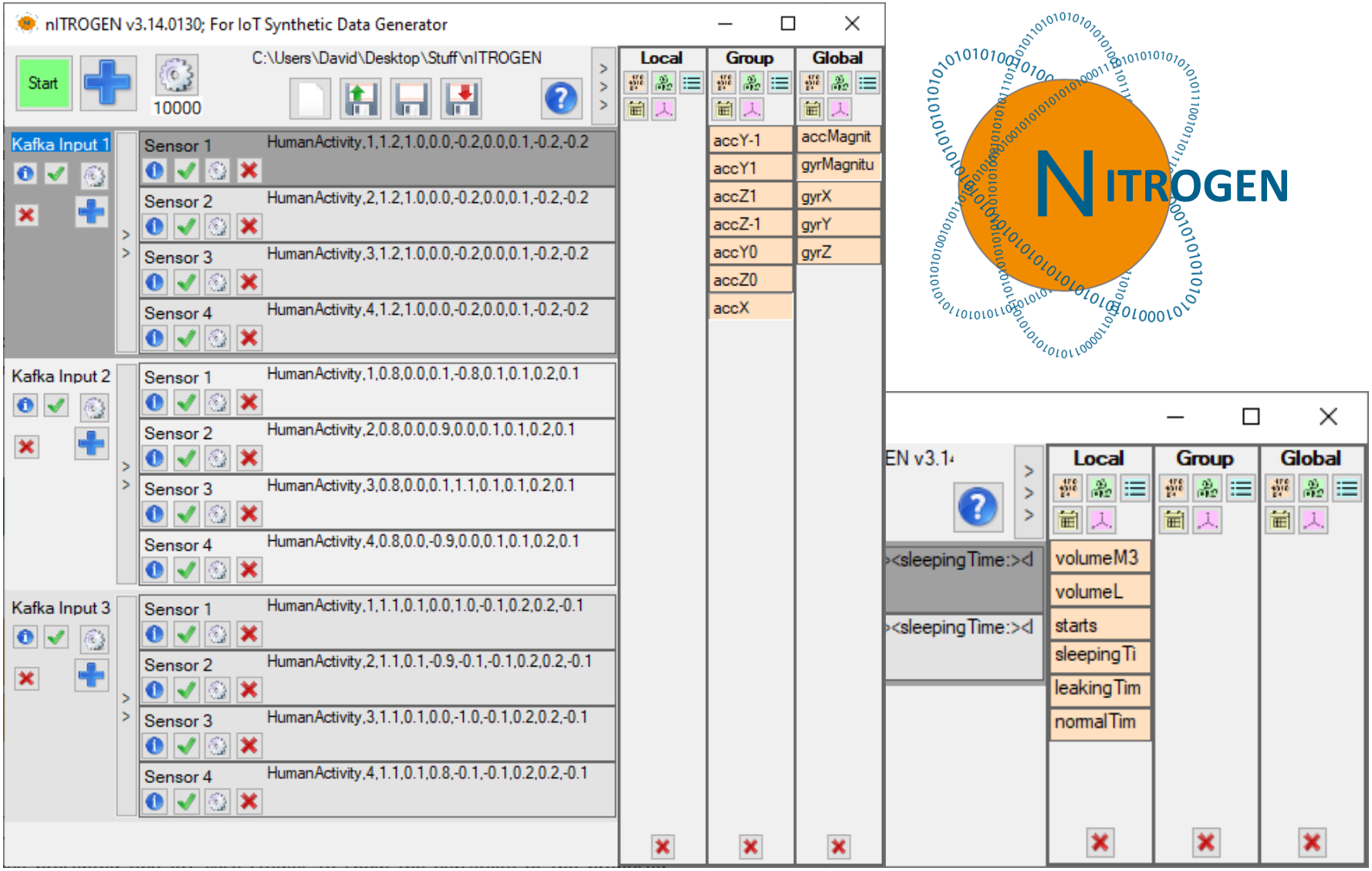
Deployed on the **SP architecture** and stored into the **NoSQL database**.



Step 2: Situation of Interest Definition



Step 3: Heterogeneous Data Simulation



Step 4: Situations of Interest Detection

Kafka Streams Case Study

Detected Alerts

Human Activity
Position: Standing or Sitting
Timestamp: 1593423368565

Human Activity
Position: Laying on back
Timestamp: 1593423368565

Human Activity
Position: Laying on stomach
Timestamp: 1593423368565

Water meter without consume
S.N.: A87SAA77188
Date: 2018-07-06T05:00:00.000Z
Volume (m3): 0
Volume (l): 6040
Type: INDUSTRIAL
Starts: 840
Batery Level: 3
Sleeping Time: 2
Leaking Time: 0
Normal Time: 6

Water meter with leaks
S.N.: A87SAA77188
Date: 2018-07-06T05:00:00.000Z
Volume (m3): 21.6
Volume (l): 218.8
Type: INDUSTRIAL
Starts: 580
Batery Level: 3
Sleeping Time: 7
Leaking Time: 2
Normal Time: 8

Water meter with wrong data
S.N.: A87SAA77188
Date: 2018-07-06T05:00:00.000Z
Volume (m3): 68.3
Volume (l): 8653.6
Type: INDUSTRIAL
Starts: 10
Batery Level: 3
Sleeping Time: 0
Leaking Time: -1
Normal Time: 2

Table of Contents

Stream Processing Architecture Performance

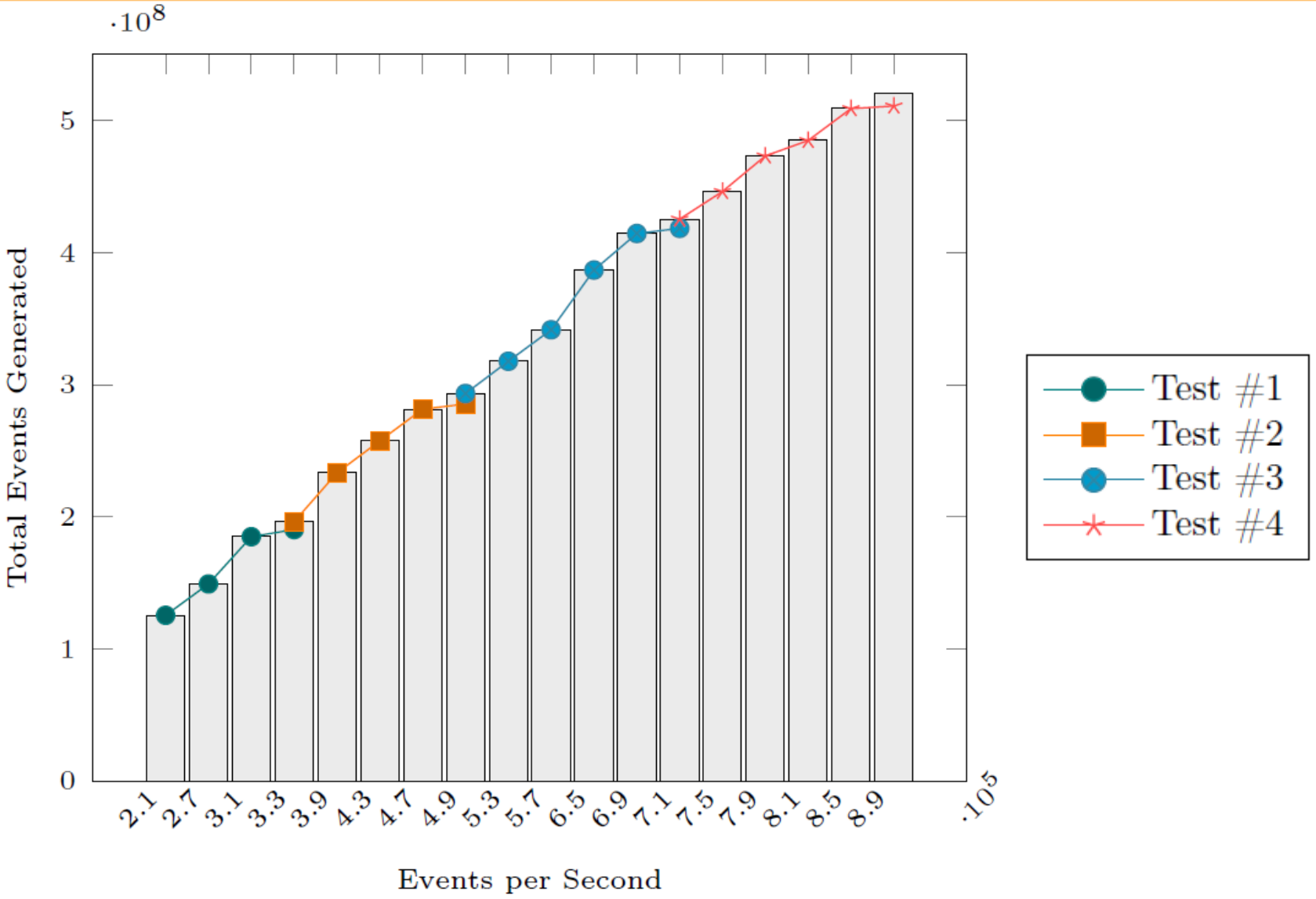
Desktop computers: Windows 10, CPU i7-4470, 12GB RAM.

JMeter and its plugins for the stress tests.

Scalability and **performance** tested in four test scenarios simulating 10-minutes tests.

- 8 partitions within a single computer.
- 16 partitions within two computers.
- 24 partitions within three computers.
- 32 partitions within four computers.

Stream Processing Architecture Performance



Stream Processing Architecture Performance

A **60-minutes** test with 150 000 e/s was also performed successfully.

Task name	Mean time (nanoseconds)
Data Homogenisation	8 183 ns
Schema generation	12 966 ns
Creation of the Avro event	2 863 ns

0.011 ms required to process each message.

100 GB of heterogeneous information processed within 10 minutes (**135 080 Mb/s**).

Table of Contents

MEdit4CEP-SP Editor Usability

A case study evaluation with **real users**.

21 students (non-experienced users) and **7 professors and researchers** (familiarized users).

They were asked to complete a case study regarding **smart water management** using **MEdit4CEP-SP**.

After complete the required tasks, they were asked to answer a **questionnaire** of 18 questions.

MEdit4CEP-SP Editor Usability

61% **prefer to use the editor** to define domains and patterns, rather than code them manually.

82% believe that **the domain expert** is the most suitable user for the editor.

100% **All** of them agreed that all the **functionalities** are **useful**.

97% were able to complete the tasks in **less than 60 minutes**, without any prior training.

75% were able to **understand** perfectly the overall **architecture**.

39% A few **improvements** were suggested.

Table of Contents

Key Features

- (F1) Scalability
- (F2) Low latency
- (F3) Real time analytics
- (F4) Data Storage
- (F5) Heterogeneous data
- (F6) Predictive Tools
- (F7) Graphical domain and alert definitions
- (F8) Runtime actions
- (F9) Persistence of definitions
- (F10) Performance evaluation

Comparative Analysis

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Our proposal	**X**	**X**	**X**	**X**	**X**	**-**	**X**	**X**	**X**	**X**
Carcillo et al.	X	-	-	X	-	X	-	-	-	X
D'Silva et al.	X	-	X	X	X	-	-	-	-	X
Jung et al.	X	-	X	-	-	-	-	-	-	X
Zeydan et al.	X	-	X	X	X	X	-	-	-	-
Malek et al.	X	-	X	X	-	-	-	-	-	X
Stripelis et al.	X	-	X	X	X	X	-	-	-	-
Hu et al.	X	-	X	-	X	-	-	-	-	-
Montori et al.	-	-	-	X	-	X	-	-	-	-
Oteafy	X	-	-	X	X	-	-	-	-	-
Santos et al.	-	-	X	X	-	X	-	-	-	-
Estévez-Ayres et al.	X	X	X	X	-	-	-	-	-	X
Boubeta et al.	X	-	X	X	-	-	X	-	-	-
Clemente et al.	X	-	X	X	-	-	X	-	-	-
Guerriero et al.	X	X	X	-	-	X	X	-	-	-

Table of Contents

Conclusions

In this PhD we have presented **MEdit4CEP-SP**.

A series of **case studies** that show its **domain versatility**.

An exhaustive evaluation that proves its **optimal performance**, its **scalability**, and its **high usability**.

A **collaborative environment** in which several users can **deploy**, **update** and **remove** definitions from a **shared** processing engine.

Future Research Lines

Time Series Classification

- **Shapelets** analysis.
- Use trends instead of event patterns.

MEdit4CEP-SP **improvements** and **extensions**

- Add **data consumers** in the editor.
- **Connect** with new analytical architectures.

Publications

International Journals

- D. Corral-Plaza et al., 2021, «MEdit4CEP-SP: A model-driven solution to improve decision-making through user-friendly management and real-time processing of heterogeneous data streams», Knowledge-Based Systems (Q1).
- D. Corral-Plaza et al., 2020, «A stream processing architecture for heterogeneous data sources in the Internet of Things», Computer Standards & Interfaces (Q1).

International Conferences

- D. Corral-Plaza et al., 2019, «A Sensor Fusion System Identifying Complex Events For Localisation Estimation», ICAC.
- L. La Blunda et al., 2019, «Distributed Real-Time Based Human Activity Analysis System», ICAC.
- D. Corral-Plaza et al., 2017, «Paving the Way for a Real-Time Context-Aware Predictive Architecture», ICSOC.

National Conferences

- D. Corral-Plaza et al., 2018, «Hacia una arquitectura para el procesamiento y análisis en tiempo real de datos heterogéneos en IoT», JCIS.
- D. Corral-Plaza et al., 2017, «Sistema para la monitorización y alerta de la calidad de aire en tiempo real», JORPRESI.

International PhD Thesis

Processing Heterogeneous Data in the Internet of Things

Thanks for your attention!

David Corral Plaza

Supervisors:
Guadalupe Ortiz Bellot, Inmaculada Medina Bulo

University of Cádiz, 8th March 2021